Theor Appl Genet (1999) 99:1031–1038

R.L. Wu · D.M. O'Malley · S.E. McKeand

Understanding the genetic architecture of a quantitative trait in gymnosperms by genotyping haploid megagametophytes

Received: 15 January 1999 / Accepted: 12 March 1999

Abstract The gymnosperms are a group of plants characterized by a haploid female gametophyte (megagametophyte). With the function of bearing the female gametes and nourishing the developing embryo, the megagametophyte has provided a simple way to understand the genetics of gymnosperm species using biochemical or genetic markers. In this paper, a quantitative genetic approach is proposed to study the genetic architecture of a quantitative trait in gymnosperms by taking advantage of the megagametophyte and the concept of average effect of a gene. Average effect describes the value associated with an allele carried by an individual and transmitted to its offspring. Through the genetic dissection of the average effect and genetic variance associated with a gamete carrying candidate genes, this approach can provide estimates of basic population genetic parameters, such as additive, dominant and epistatic effects, allelic frequencies and linkage disequilibrium. The candidate genes, known through their major mutant phenotype, have been reported in gymnosperms. An example for a candidate gene affecting lignin biosynthesis was applied to demonstrate the statistical procedures of the approach and its advantage. The conditions upon which the approach can be effectively used are discussed.

Key words Average effect · Candidate gene · Genetic architecture · Gymnosperm · Megagametophyte

Communicated by P.M.A. Tigerstedt

R.L. Wu (✉) · D.M. O'Malley · S.E. McKeand
Forest Biotechnology Group,
Department of Forestry,
North Carolina State University,
Raleigh, NC 27695–8008, USA
e-mail: rwu@statgen.ncsu.edu
Fax: +1 919 515 7135

Present address:
R.L. Wu
Program in Statistical Genetics,
Department of Statistics,
Box 8203, North Carolina State University,
Raleigh, NC 27695-8203, USA

Introduction

The gymnosperms, a group of plants with a well-documented evolutionary history, originated over 300 million years ago, before the flowering plants (Bierhorst 1971). Today, they are of tremendous economic importance for lumber and wood pulp and are represented by such familiar trees as pine, fir, spruce, cedar, redwood, yew and ginkgo. Modern gymnosperms have a worldwide distribution and in temperate zones are the dominant trees, forming vast forests in many regions. Gymnosperms have the characteristic of naked seeds, which means that the seeds are usually borne on the surface of scales rather than enclosed in fruits as in the flowering plants. Gymnosperms are characterized by a haploid female gametophyte (megagametophyte, 1*n*) that serves a nutritive tissue surrounding the embryo in a mature seed.

Being haploid, the megagametophyte is excellent genetic material to characterize individual alleles using biochemical or genetic technologies. Genetic analysis of the megagametophyte using isozymes has been carried out in conifers for many years for the estimation of genetic diversity, heterozygosity, genetic relatedness and for studies of gene flow in natural populations (Wheeler and Guries 1982; Millar 1983; Hamrick et al. 1992; Huang et al. 1994; Wang et al. 1996; Rogers 1997; Wang and Nagasaka 1997; Wang and Liu 1998). More recently, attempts have been made to employ the megagametophyte to construct genetic linkage maps directly based on open-pollinated progenies from a single heterozygous tree using polymerase chain reaction (PCR)-based dominant markers. The species mapped include *Pinus taeda* (J.J. Mackay 1996; Wilcox et al. 1996), *Pinus pinaster* (Plomion et al. 1995), *Pinus elliottii* (Nelson et al. 1993), *Pinus palustris* (Nelson et al. 1994), *Picea glauca* (Tulserium et al. 1992) and *Picea abies* (Binelli and Bucci 1994). Potentially, these maps can offer detailed studies of forest tree genome structure and function (Neale and Williams 1991) and of the genetic dissection of complex traits (e.g. Plomion et al. 1996).

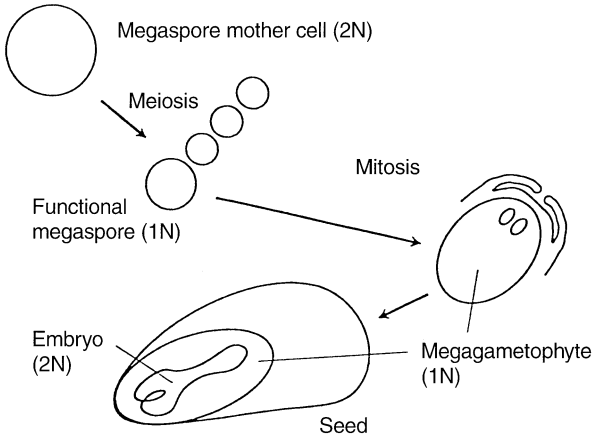

Fig. 1 The biology of haploid megagametophyte. Development of the megagametophyte proceeds from a single haploid product of meiosis. The functional megaspore divides by mitosis to produce the megagametophyte, which will develop the archegona that contain the female gametes. Several gametes could be fertilized by pollen but only ones give rise to the embryo in the mature seed

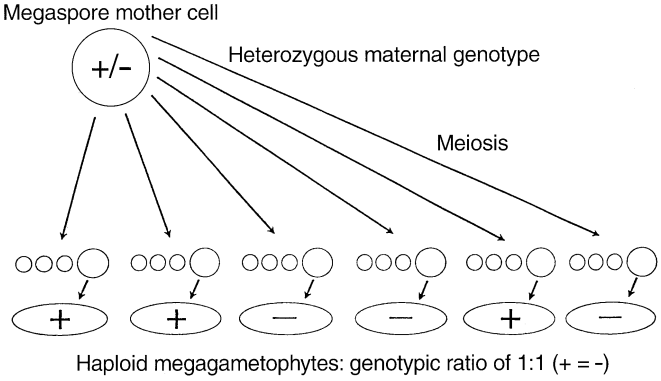

Fig. 2 The megagametophyte in each mature seed is genetically equivalent to a haploid progeny plant, thus any heterozygous locus in the seed parent will segregate 1:1 in the megagametophytes, regardless of the pollen contribution

The megagametophyte also provides a simple avenue by which to study the relationship between markers and phenotypic traits at the population level based on the concept of average effect of a gene. The average effect of a particular allele is the deviation from the population mean of individuals which receive that allele from one parent and the second allele from the other parent randomly sampled from the population (Falconer and Mackay 1996). Thus, by genotyping the megagametophyte, the average effect of an allele segregating in a natural population can be readily estimated. Average effects depend on genotypic values and allelic frequencies, and their genetic dissection provides a unique power for estimating these parameters for the segregating gene. In this paper, we extend this principle to estimate epistatic effects between two different genes of known function and chromosomal positions.

Megagametophyte biology and experimental design

The megagametophyte of gymnosperms is a multicelluar structure and serves the double function of bearing the gametes as well as nourishing the developing embryo (Maheshwari and Singh 1967). The morphology and development of the megagametophyte have been used as a criterion for plant taxonomy (Linder and Rudall 1993). The formation of megagametophyte is initiated through the meiotic division of the diploid megaspore mother cell in which four haploid meiotic products, megaspores, are produced (Fig. 1). Three megaspores apparently degenerate while the fourth, functional megaspore, divides by mitosis to produce the megagametophyte, which will develop the archegonia that contain the female gametes. The allelic contribution of the seed parent to the embryo segregates in megagametophytes from that tree (Fig. 2). The genotype of each megagametophyte is identical to that of the maternal gamete that forms the embryo. The megagametophyte in each seed is genetically equivalent to a haploid progeny plant, therefore any heterozygous locus in the seed parent will segregate 1:1 in the megagametophytes, regardless of the pollen contribution (Fig. 2).

Consider a tree that is heterozygous for alleles of candidate genes of interest. Seeds collected from the heterozygous tree are germinated in the greenhouse. While seedlings are transplanted to a field trial, megagametophytes are collected for the isolation of genomic DNA. Because the megagametophyte is haploid, PCR-based dominant markers, such as random amplified polymorphic DNA (RAPDs) and amplified fragment length polymorphisms (AFLPs), can be effectively used to genotype each megagametophyte by scoring the presence or absence of bands. The field trial for seedlings is established in a complete randomized design. For some coniferous species that are easy to propagate vegetatively, such as *Abies*, *Larix* and *Picea*, the use of clonal replicates is suggested because this can significantly increase within-family heritability and the precision of progeny testing. Quantitative traits, such as growth, crown architecture, wood properties and physiological processes, are measured for each tree.

Model

A single gene

Consider a heterozygous tree in a random mating population. If the species considered is monoecious, this tree may be pollinated simultaneously by its own pollen and other unrelated trees' pollen. As a result, seeds collected from the mother tree include those from both selfed and outcrossed pollination. For the selfed pollination, maternal and paternal gametes at a genetic locus may be assumed to have the identical frequency because they are derived from the same genotype. However, for the outcrossed progeny, although maternal gametes are the same as those for the selfed progeny, paternal gametes come from the natural population (excluding the mother tree), and their frequency is thus are determined by the population structure.

Assume that there is only one candidate gene on a linkage group. This candidate gene has two alleles, A_1 and A_2, characterized from megagametophytes derived from the heterozygous tree. The population frequencies of these two alleles are denoted by p and q, respectively. Each of these two alleles expressed in the maternal gametes has been randomly united with alleles either from the same tree or the population to form embryos. If the outcrossing rate is t, then the probability for A_1 to unite

Table 1 Estimate of the average effects of a gene using the marker analysis of megagametophytes derived from a heterozygeous tree in the gymnosperm population

Maternal gamete	Values and frequency of embryo genotypes			Mean values of Embryo genotypes	Average effect of Gene
	A_1A_1 a	A_1A_2 d	A_2A_2 $-a$		
A_1	$tp+\frac{1}{2}(1-t)$	$tq+\frac{1}{2}(1-t)$		$t(ap+dq)+\frac{1}{2}(1-t)(a+d)$	$\alpha_{A_1}=\frac{1}{2}a+d(q-\frac{1}{2})t$
A_2		$tp+\frac{1}{2}(1-t)$	$tq+\frac{1}{2}(1-t)$	$-t(aq-dp)-\frac{1}{2}(1-t)(a-d)$	$\alpha_{A_2}=-\frac{1}{2}a+d(p-\frac{1}{2})t$

with A_1 and A_2 is $tq+\frac{1}{2}(1-t)$ and $tp+\frac{1}{2}(1-t)$, respectively. Similar expressions can be obtained for A_2 (Table 1). Because the candidate gene exerts direct effects on a quantitative trait, its additive (a) and dominant effects (d) can be estimated based on its segregation pattern in a progeny population. The average effect of an allele (α_{A1} or $\alpha_{A2)}$ is calculated by subtracting the population mean, $\mu=\frac{1}{2}a(p-q)t+\frac{1}{2}d$, from the mean value of genotypes produced by this allele (Table 1):

$$\alpha_{A_1} = \tfrac{1}{2}a + d\left(q-\tfrac{1}{2}\right)t \quad (1a)$$

$$\alpha_{A_2} = -\tfrac{1}{2}a + d\left(p-\tfrac{1}{2}\right)t \quad (1b)$$

The average effect of the gene substitution is the difference between the average effects of the two alleles,

$$\alpha = \alpha_{A_1} - \alpha_{A_2} = a + d(q-p)t \quad (1c)$$

From Table 1, the genetic variance within each of the two megagametophyte genotypes, A_1 or A_2, can be derived as:

$$V_{A_1} = \tfrac{1}{4}[1+(q-p)t][1+(p-q)t](a-d)^2 \quad (2a)$$

$$V_{A_2} = \tfrac{1}{4}[1+(q-p)t][1+(p-q)t](a+d)^2 \quad (2b)$$

The sum (S) and difference of these two variances (Δ) are:

$$S = \tfrac{1}{2}[1+(q-p)t][1+(p-q)t(a^2+d^2)] \quad (2c)$$

$$\Delta = -ad[1+(q-p)t][1+(p-q)t] \quad (2d)$$

If a field trial using the progeny of the mother tree includes clonal replicates for each individual, then $V_{A_1}$and V_{A_1} can be estimated accurately by a simple analysis of variance. Thus, Eqs. 2a and 2b represent two independent equations and their simplified versions, Eqs. 2c and 2d, along with Eq. 1c, construct a system of three equations to obtain the solutions of a and d by solving the nonlinear equations (MATHEMATICA, Wolfram 1996):

$$\Delta a^2 + \Delta d^2 + 2Sad = 0 \quad (3a)$$

$$ad^2 + \Delta d - a(\alpha - a)^2 = 0 \quad (3b)$$

Estimates of a and d are further used to estimate the allelic frequency of A_1 or A_2 in the natural population. It is not difficult to derive two different expressions of estimating the difference of allelic frequencies:

$$q-p = \begin{cases} \dfrac{\alpha - a}{dt} & \text{when } a = d \\ \pm t\sqrt{\dfrac{(a-d)^2 - 4V_{A_1}}{(a-d)^2}} & \text{when } d = 0 \end{cases} \quad (3c)$$

When $a \neq d$ and $d \neq 0$, either expression in Eq. 3C can provide the estimate for $g-p$. The outcrossing rate of a population can be estimated easily by using molecular markers, such as RAPDs, RFLPs or microsatellites (e.g. Burczyk et al. 1997; Zheng and Ennos 1997; Wang and Liu 1998).

In a situation where clones are unavailable, estimates of V_{A_1} and V_{A_2} each include a residual variance due to environmental errors and other linked gene effects. It is not unreasonable to assume that the residual variance is the same between the two allele groups. Thus, such a residual variance can be eliminated by taking Eq. 2d, which constructs a system of two equations with Eq. 1c, to solve a and d:

$$a = \tfrac{1}{2}\left[\alpha \pm \sqrt{\alpha^2 + \frac{4(q-p)t}{[1+(q-p)t][1-(q-p)t]}\,\Delta}\right] \quad (4a)$$

$$d = \begin{cases} \dfrac{\alpha - a}{(q-p)t} & \text{when } a = 0 \\ -\dfrac{\Delta}{[1+(q-p)t][1-(q-p)t]} & \text{when } q = p \text{ or } t = 0 \end{cases} \quad (4b)$$

Allelic frequency cannot be estimated due to inadequate degrees of freedom, but its value interval can be given as follows:

$$\frac{2\Delta - \sqrt{4\Delta^2 + \alpha^4}}{\alpha^2 t} \leq q-p \leq \frac{2\Delta + \sqrt{4\Delta^2 + \alpha^4}}{\alpha^2 t} \quad (4c)$$

However, only when the left term of the inequality is greater than −1 and only when the right term is less than 1 would this constraint be useful to determine the finer interval of allelic frequency. After the interval of allelic frequency is determined, the values for a and d can be approximated by a numeric simulation on the interval.

Two unlinked, epistatically-interacting genes

The idea described in the preceding section can be extended to estimate epistatic effects between two candidate genes of known biological functions. It is still assumed that there is only a single candidate gene on a linkage group. Consider two candidate genes that are on different linkage groups that are in linkage equilibrium but epistatically interact to affect a quantitative trait in a population of gymnosperm. The alleles of one gene are denoted by A_1 and A_2 and the alleles of the other gene by

Table 2 Estimate of the average effects of maternal gametes at two unlinked but epistatically interacting genes using the marker analysis of megagametophytes in gymnosperm

Maternal gamete/frequency	Embryo genotype and its value and frequency				Mean value of embryo genotype	Average effect of gene
A_1B_1 $\frac{1}{4}$	$A_1A_1B_1B_1$ $a+a'+i$ $tpp'+\frac{1}{4}(1-t)$	$A_1A_1B_1B_2$ $a+d'+j$ $tpq'+\frac{1}{4}(1-t)$	$A_1A_2B_1B_1$ $d+a'+j$ $tqp'+\frac{1}{4}(1-t)$	$A_1A_2B_1B_2$ $d+d'+l$ $tqq'+\frac{1}{4}(1-t)$	$\mu_{A_1B_1}$	$\alpha_{A_1B_1}$
A_1B_2 $\frac{1}{4}$	$A_1A_1B_1B_2$ $a+d'+j$ $tpp'+\frac{1}{4}(1-t)$	$A_1A_1B_2B_2$ $a-a'-i$ $tpq'+\frac{1}{4}(1-t)$	$A_1A_2B_1B_2$ $d+d'+l$ $tqp'+\frac{1}{4}(1-t)$	$A_1A_2B_2B_2$ $d-a'-j$ $tqq'+\frac{1}{4}(1-t)$	$\mu_{A_1B_2}$	$\alpha_{A_1B_2}$
A_2B_1 $\frac{1}{4}$	$A_1A_2B_1B_1$ $d+a'+j$ $tpp'+\frac{1}{4}(1-t)$	$A_1A_2B_1B_2$ $d+d'+l$ $tpq'+\frac{1}{4}(1-t)$	$A_2A_2B_1B_1$ $-a+a'-i$ $tqp'+\frac{1}{4}(1-t)$	$A_2A_2B_1B_2$ $-a+d'-j$ $tqq'+\frac{1}{4}(1-t)$	$\mu_{A_2B_1}$	$\alpha_{A_2B_1}$
A_2B_2 $\frac{1}{4}$	$A_1A_2B_1B_2$ $d+d'+$l $tpp'+\frac{1}{4}(1-t)$	$A_1A_2B_2B_2$ $d-a'-j$ $tpq'+\frac{1}{4}(1-t)$	$A_2A_2B_1B_2$ $-a+d'-j$ $tqp'+\frac{1}{4}(1-t)$	$A_2A_2B_2B_2$ $-a-a'+I$ $tqq'+\frac{1}{4}(1-t)$	$\mu_{A_2B_2}$	$\alpha_{A_2B_2}$

B_1 and B_2. Additive and dominant effects at gene **A** are a and d, whereas those at gene **B** are a' and d'. Epistatic effects between the two genes are i for the additive×additive interaction, j for the additive×dominant or dominant×additive interaction and l for the dominant×dominant interaction. Four maternal gametes from a heterozygous tree are randomly combined to paternal gametes from the same tree or other trees in the population. Haploid megagametophytes derived from the mother tree are genotyped at these two genes by using a molecular marker technology. Assuming that the population frequencies of the alleles are p for A_1 and q for A_2 and p' for B_1 and q' for B_2, respectively, and that the outcrossing rate is t, average effects of the four maternal gametes, A_1B_1, A_1B_2, A_2B_1 and A_2B_2, identified from the megagametophytes can be estimated (Table 2). The mean value of a quantitative trait corresponding to each megagametophyte genotype can be expressed by:

$$\mu_{A_1B_1} = t[pa + qd + p'a' + q'd' + pp'i + (pq' + qp')j + qq'l] + \tfrac{1}{2}(1-t)\left(a + a' + d + d' + \tfrac{1}{2}i + j + \tfrac{1}{2}l\right) \quad (5a)$$

$$\mu_{A_1B_2} = t[pa + qd - q'a' + p'd' - pq'i + (pp' - qq')j + qp'l] + \tfrac{1}{2}(1-t)\left(a - a' + d + d' - \tfrac{1}{2}i + \tfrac{1}{2}l\right) \quad (5b)$$

$$\mu_{A_2B_1} = t[-qa + pd + p'a' + q'd' - qp'i + (pp' - qq')j + qp'l] + \tfrac{1}{2}(1-t)\left(-a + a' + d + d' - \tfrac{1}{2}i + \tfrac{1}{2}l\right) \quad (5c)$$

$$\mu_{A_2B_2} = t[-qa + pd - q'a' + p'd' + qq'i - (pq' + qp')j + pp'l] + \tfrac{1}{2}(1-t)\left(-a - a' + d + d' + \tfrac{1}{2}i - j + \tfrac{1}{2}l\right) \quad (5b)$$

The average effect of a megagametophyte genotype is the difference between its mean value and the overall mean, $\mu = \frac{1}{2}\{t[(p-q)a + d + (p'-q')a' + d' + \frac{1}{2}(p-q)(p'-q')i + (pp' - qq')j + \frac{1}{2}l] + (1-t)(d + d' + \frac{1}{2}l)\}$. The equations for estimating the average effect as well as the genetic variance of each of the four megagametophyte genotypes are given in Wu (1998). The genetic variance of average effects across the four megagametophytes at the two candidate genes can be estimated by:

$$V = \tfrac{1}{4}\left[\alpha^2_{A_1B_1} + \alpha^2_{A_1B_2} + \alpha^2_{A_2B_1} + \alpha^2_{A_2B_2}\right] \quad (6)$$

If genetic variance within each of the four megagametophyte genotypes can be estimated accurately, e.g. using clonal replicates, one can construct a system of nine independent equations which include the average effect and genetic variance of each of the four megagametophyte genotypes and Eq. 6. These nine equations are used to solve nine unknown parameters, a, d, a', d', i, j, l, p and p', using MATHEMATICA (Wolfram 1996). Outcrossing rate is estimated independently using molecular markers.

If genetic variances within the megagametophyte genotypes cannot be estimated accurately, one should take the differences between the genetic variances within different megagametophyte genotypes to eliminate residual variances included in each megagametophyte genotype (see above). In this case, the number of equations is reduced to eight, which are not adequate to obtain the solutions of the nine unknowns. However, if we pose some restrictions on at least one of the unknowns, approximate solutions can be obtained. Because allelic frequencies p and p' must be a value between 0 and 1, restrictions on these two parameters should be the first choice. If p and p' are fixed, the number of equations (eight) will be greater than the number of unknowns to be estimated (seven). In this case, a nonlinear optimization approach based on the weighted least squares analysis can be used to obtain the approximation estimate for each variable and its sampling error. The weights for the differences of across-gamete average effect and genetic variance are determined by the reciprocals of their sampling variances multiplied by the corresponding degrees of freedom. The use of these weights results in the solutions of the unknowns and their standard errors by the nonlinear optimization manipulation. These afford the means of calculating the weights for the second round of the iterative procedure. This step is repeated until the final estimates converge to stable values. The adequacy of the model, i.e. the degree to which esti-

Table 3 Estimate of the average effects of maternal gametes at two linked and epistatically interacting gene using the marker analysis of megagametophytes in gymnosperm

Maternal gamete/frequency	Embryo genotype and its value and frequency				Mean value of embryo genotype	Average effect of gene
A_1B_1 $\frac{1}{2}(1-r)$	$A_1A_1B_1B_1$ $a+a'+i$ $t(pp'+D)+$ $\frac{1}{2}(1-t)(1-r)$	$A_1A_1B_1B_2$ $a+d'+j$ $t(pq'-D)+$ $\frac{1}{2}(1-t)(1-r)$	$A_1A_2B_1B_1$ $d+a'+j$ $t(qp'-D)+$ $\frac{1}{2}(1-t)(1-r)$	$A_1A_2B_1B_2$ $d+d'+l$ $t(qq'+D)+$ $\frac{1}{2}(1-t)(1-r)$	$\mu_{A_1B_2}$	$\alpha_{A_1B_2}$
A_1B_2 $\frac{1}{2}r$	$A_1A_1B_1B_2$ $a+d'+j$ $t(pp'+D)+\frac{1}{2}(1-t)r$	$A_1A_1B_2B_2$ $a-a'-i$ $t(pq'-D)+\frac{1}{2}(1-t)r$	$A_1A_2B_1B_2$ $d+d'+l$ $t(qp'-D)+\frac{1}{2}(1-t)r$	$A_1A_2B_2B_2$ $d-a'-j$ $t(qq'+D)+\frac{1}{2}(1-t)r$	$\mu_{A_1B_2}$	$\alpha_{A_1B_2}$
A_2B_1 $\frac{1}{2}r$	$A_1A_2B_1B_1$ $d+a'+j$ $t(pp'+D)+\frac{1}{2}(1-t)r$	$A_1A_2B_1B_2$ $d+d'+l$ $t(pq'-D)+\frac{1}{2}(1-t)r$	$A_2A_2B_1B_1$ $-a+a'-i$ $t(qp'-D)+\frac{1}{2}(1-t)r$	$A_2A_2B_1B_2$ $-a+d'-j$ $t(qq'+D)+\frac{1}{2}(1-t)r$	$\mu_{A_2B_1}$	$\alpha_{A_2B_1}$
A_2B_2 $\frac{1}{2}(1-r)$	$A_1A_2B_1B_2$ $d+d'+l$ $t(pp'+D)+$ $\frac{1}{2}(1-t)(1-r)$	$A_1A_2B_2B_2$ $d-a'-j$ $t(pq'-D)+$ $\frac{1}{2}(1-t)(1-r)$	$A_2A_2B_1B_2$ $-a+d'-j$ $t(qp'-D)+$ $\frac{1}{2}(1-t)(1-r)$	$A_2A_2B_2B_2$ $-a-a'+i$ $t(qq'+D)+$ $\frac{1}{2}(1-t)(1-r)$	$\mu_{A_2B_2}$	$\alpha_{A_2B_2}$

mates of the unknowns fit the eight equations simultaneously, is tested based on the sum of squares of differences between the expected and observed values for the left terms of these equation (each square being multiplied by the weight) using χ^2-statistics. The degrees of freedom for the χ^2-test of goodness-of-fit would be the number of nonlinear equations used in the model minus the number of variables to be estimated. The adequacy of the model can be evaluated by the probability (P) level, which is the probability of getting a χ^2-value larger than the value actually obtained, given that the fixed values of p and p' are correct. The small values of P correspond to a poor fit and large values to a good fit. When the model has the best adequacy, the estimators of unknowns and the fixed values of p and p' are considered as the closest to their actual values.

Two linked, epistatically-interacting genes

If two epistatically interacting candidate genes of known biological functions are on the same linkage group, they will likely be in linkage disequilibrium in the population. The population frequency of each paternal gamete at genes **A** and **B** is a function of allelic frequency at each gene and the gametic-phase linkage disequilibrium between these two genes, D. The mean value of a quantitative trait corresponding to each megagametophyte genotype, A_1B_1, A_1B_2, A_2B_1 or A_2B_2, can be derived from Table 3:

$$\begin{aligned}\mu_{A_1B_1} &= t[pa+qd+p'a'+q'd'+(pp'+D)i \\ &\quad +(pq'+qp'-2D)j+(qq'+D)l]+(1-t)(1-r) \\ &\quad \cdot\left(a+a'+d+d'+\tfrac{1}{2}i+j+\tfrac{1}{2}l\right) \qquad (7a)\\ \mu_{A_1B_2} &= t[pa+qd-q'a'+p'd'-(pq'-D)i \\ &\quad +(pp'-qq')j+(qp'-D)l]+(1-t)(1-r) \\ &\quad \cdot\left(a-a'+d+d'-\tfrac{1}{2}i+\tfrac{1}{2}l\right) \qquad (7b)\\ \mu_{A_2B_1} &= t[-qa+pd+p'a'+q'd'-(qp'-D)i \\ &\quad +(pp'-qq')j+(pq'-D)l]+(1-t)(1-r) \\ &\quad \cdot\left(-a+a'+d+d'-\tfrac{1}{2}i+\tfrac{1}{2}l\right) \qquad (7c)\\ \mu_{A_2B_2} &= t[-qa+pd-q'a'+p'd'+(qq'+D)i \\ &\quad -(pq'+qp'-2D)j+(pp'+D)l]+(1-t)(1-r) \\ &\quad \cdot\left(-a-a'+d+d'+\tfrac{1}{2}i-j+\tfrac{1}{2}l\right), \qquad (7d)\end{aligned}$$

where r is the recombination frequency between the two candidate genes, which can be estimated based on their segregation in megagametophytes, and the other parameters are as defined in the preceding section. The average effect of each megagametophyte genotype is the difference between its mean value and the overall mean, $\mu=\frac{1}{2}\{t(p-q)a+d+(p'-q')a'+d'+(pp'+qq'+2D)i+r(pp'-qq')j+\frac{1}{2}[pp'+qq'+2D+r(pq'+qp'-pp'-qq'-4D)l]\}+(1-t)(1-r)(d+d'+\frac{1}{2}i-ri+l)$. The equations for estimating average effects and genetic variances of the four megagametophyte genotypes at two linked, epistatically interacting genes are given in Wu (1998). The genetic variance of average effects at the two genes can be estimated by:

$$V=\frac{1}{1-r}\left[\alpha^2_{A_1B_1}+\alpha^2_{A_2B_2}\right]+\frac{1}{r}\left[\alpha^2_{A_1B_2}+\alpha^2_{A_2B_1}\right] \qquad (8)$$

Because there are ten unknown parameters for two linked, epistatically interacting genes (nine of them are the same as for two unlinked epistatic genes and the tenth is the gametic-phase linkage disequilibrium, D), a system of nine independent equations are not adequate to obtain the solutions of the unknowns. By setting p and p' to be fixed, however, a nonlinear optimization procedure, as described above, can be used to obtain an approximate solution of each unknown. Owing to more parameters being estimated for two linked epistatic genes, the use of clonal replicates is very important for increasing the number of independent equations by obtaining the accurate estimate of the genetic variance within each megagametophyte genotype.

Example

Lignin is a complex phenolic polymer that reinforces the walls of certain cells in the vascular tissues of higher plants (Sederoff et al. 1994). Lignin plays a very important role in maintaining mechanical support, water transportation and defense against diseases and pests. However, during the pulping process, lignin must be removed at both an environmental and an economical cost to make good quality paper. Consequently, genetic approaches have been proposed to modify lignin content to a level, at which plants could grow normally but the cost of its removal would be minimum (Campbell and Sederoff 1996).

Recently, a mutant allele of the *cad* gene affecting lignification was discovered in a woody plant by J.J. Mackay (1996). *cad* is a gene that encodes for the monolignol biosynthetic enzyme cinnamyl alcohol dehydrogenase (CAD, E.C. 1.1.1.195), which catalyzes the final step of lignin precursor biosynthesis, the reduction of cinnamaldehydes to cinnamyl alcohols. The mutated allele, named *cad-nl*, was further found to be favorably associated with stem growth in loblolly pine (Wu et al. 1999). The *cad* gene with a known function and position can be viewed as a candidate gene. Based on its segregation in haploid megagametophytes from a heterozygous tree, the additive and dominant effects of *cad* on growth traits are estimated using the model proposed in this paper.

Plant material used in this example was derived from open-pollinated progenies from an offspring (identified as selection 7–1037) of a cross between loblolly pine genotypes 7–56, an original tree from which mutation in the *cad* gene was discovered, and 7–51. In 1993, 900 seedlings derived from the seeds of selection 7–1037 were transplanted to a field trial with nine square blocks in Lumberton, N.C. Height growth was measured for each tree in the plantation at the end of each of the first 2 years from which second-year shoot elongation was calculated. Megagametophytes were collected following seedling germination and used to genotype *cad* genotypes using a pair of 20-bp custom primers designed from the DNA sequence of the region of the *cad* gene from genotype 7–56. In this example, we would like to emphasize the statistical procedures of estimating epistasis between two known genes by genotyping megagametophytes. We used a RAPD marker, amplified by primer H15 and associated significantly with growth traits in the open-pollinated progeny of selection 7–1037 (R.L. Wu, unpublished data), as the surrogate of the second candidate gene. This RAPD marker, named H15_750, is not linked with *cad*.

Average effects and phenotypic variances associated with four megagametophyte genotypes for *cad* and H15_750 were calculated for second-year shoot elongation (in the unit of centimeters). Because no clonal replicates were used, the differences between the phenotypic variances associated with these megagametophyte genotypes were calculated to eliminate the influences of residual variances. There have been many observations on outcrossing rate for coniferous populations, which, for example, suggest t=0.80 to 0.90 for *Pinus caribaea* (Zheng and Ennos 1997) and t=0.89 to 0.97 for *Pinus attenuata* (Burczyk et al. 1997). In this example, outcrossing rate is assumed to be 0.90. Let the frequency of mutant allele *cad-nl* be p for the *cad* gene and the frequency of the favorable allele be p' at H15_750. Under different combinations of p and p', additive, dominant and epistatic effects at these two loci were estimated, along with the χ^2 statistics with which the adequacy of the model is tested. It was found that a minimum χ^2 value ($\chi^2_{df=1}$=1.32, P>0.50), one at which the model has the best adequacy, was around a combination of p=0.20 and p'=0.45. Under this combination, estimates of additive and dominant effects were a=17 cm and d=18 cm at *cad*, and a'=15 cm and d'=10 cm at H15_750, respectively, and estimates of epistatic effects at these two loci were i=18 cm, j=l0 cm and l=9 cm.

Discussion

The genetics of forest trees, especially gymnosperms, is very difficult to understand because classical genetic material, such as inbred lines, is not available for these species. Although abundant natural variation exists in forest trees, its exploitation in operational forestry has been very limited with little knowledge of the genetic basis underlying it. However, gymnosperms also have a major advantage, that is the haploid nature of the megagametophyte. The megagametophyte is an excellent natural resource for characterizing individual alleles using molecular marker technologies. Its applications to studying population structure and constructing linkage maps have received great attention in major coniferous species. As demonstrated in this paper, the integration of the megagametophyte and the concept of average effect of a gene can shed light on the genetic structure of a quantitative trait in a natural population of gymnosperms.

The idea used in this paper is not complicated. It includes two key derivation steps, one based on the average effect of a gamete carrying the candidate genes of interest and the other on the genetic variance within the gamete. The traditional concept of average effect has been defined based on a given allele that is randomly combined to alleles from their population (Falconer and Mackay 1996). Although average effect is a very important concept for describing population structure and evolution, its magnitude cannot be estimated because it is associated with the value of individual genes rather than genotypes. The megagametophyte represents the genotype identical to that of the female gamete and usually contains enough DNA enough to perform marker analysis. Both properties provide the means by which to estimate the average effect of an allele in the population. This allele has been randomly united with alleles from the population to form the embryo that develops into a seedling from which the phenotypic measurement is

made. Here, we extend the concept of average effect to the level of the gamete that carries different alleles from different loci. We have also derived the genetic variances associated with these gametes. Basic genetic parameters, such as additive, dominant and epistatic effects, allelic frequencies and linkage disequilibrium, contributing to the average effect and genetic variance within a gamete are estimated by solving a group of independent nonlinear equations.

The analysis here has involved the fitting of candidate genes with known function and chromosomal positions. Candidate genes have been detected in many species (T.F.C. Mackay 1996). For example, in mice, mutations at five loci (*agouti*, *diabets*, *obese*, *tubby* and *fat*) result in grossly obese phenotypes (Chua 1997). All five genes have been cloned, and their map positions are accurately known. In gymnosperms, there have been a few reports on the detection of candidate genes. Mackay et al. (1997) identified a mutant allele, *cad-nl*, which reduces the gene expression of lignin biosynthesis by 50% as compared to its wild-type allele. In *Pinus radiata*, a homologue *needly* of the flower meristem-identity genes, *leafy*/*floricaula*, from *Arabidopsis* and *Antirrhinum* has been identified as affecting the vegetative development of pine (Mouradov et al. 1998). For a practical analysis, it would be preferable to use markers within the candidate genes themselves, but it seems reasonable to use tightly linked markers from genetic maps. However, if neither candidate genes nor tightly linked markers exist, the current analysis should be modified to detect real quantitative trait loci that affect variation in a quantitative trait with the aid of markers.

It should be pointed out that the power for estimating these genetic parameters is dependent on the use of clonal replicates for the seedlings planted. If the seedlings are cloned, the genetic variances associated with gametes can be well estimated by an analysis of variance. Otherwise, the estimated genetic variances include influences of environmental errors and other linked genes. A way to eliminate the influences is to take differences between these genetic variances by assuming that the same residual variance is included in each gamete. In this situation, because of a reduced number of independent equations, the power of the model is reduced. The analytical method described in this paper can effectively estimate epistasis between two linked or unlinked genes. However, when there are more than two genes on a single linkage group, a more complicated maximum likelihood method should be used to dissect these individual genes (Wu 1999). In practice, results obtained from the current approach may provide basic ideas about the effective use of the maximum likelihood method.

Acknowledgments We thank Prof. R.R. Sederoff and other members of the Forest Biotechnology Group at North Carolina State University for continuous support on this and other studies. We are grateful to Drs. Z.-B. Zeng and S. Xu for helpful comments on this manuscript. This work was partially supported by the NCSU Industrial Biotechnology Associates and the NCSU-Industrial Tree Improvement Cooperative.

References

Bierhorst DW (1971) Morphology of vascular plants. Macmillan, New York

Binelli G, Bucci G (1994) A genetic-linkage map of *Picea abies* Karst, based on RAPD markers, as a tool in population-genetics. Theor Appl Genet 88: 283–288

Burczyk J, Adams WT, Shimizu JY (1997) Mating system and genetic diversity in natural populations of knobcone pine (*Pinus attenuata*). For Genet 4: 223–226

Campbell MM, Sederoff RR (1996) Variation in lignin content and composition: mechanisms of control and implications for the genetic improvement of plants. Plant Physiol 110:3–13

Chua SC JR (1997) Monogenic models of obesity. Behav Genet 27:277–284

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Longman Group, Harlow, Essex

Hamrick JL, Godt MJW, Sherman-Broyles SL (1992) Factors influencing levels of genetic diversity in woody plant species. New For 6:95–124

Huang QQ, Tomaru N, Wang LH, Ohba K (1994) Genetic-control of isozyme variation in masson pine, *Pinus massoniana* Lamb. Silvae Genet 43:285–292

Linder HP, Rudall PJ (1993) The megagametophyte in anarthria (*Anarthriaceae, Poales*) and its implications for the phylogeny of the poales. Am J Bot 80:1455–1464

Mackay JJ (1996) A Mutation in lignin biosynthesis in loblolly pine: genetic, molecular and biochemical analyses. PhD thesis. North Carolina State University, Raleigh, N.C.

Mackay JJ, O'Malley DM, Presnell T, Booker FL, Campbell MM, Whetten RW, Sederoff RR (1997) Inheritance, gene expression, and lignin characterization in a mutant pine deficient in cinnamyl alcohol dehydrogenase. Proc Natl Acad Sci USA 94:8255–8260

Mackay TFC (1996) The nature of quantitative genetic variation revisited: lessons from *Drosophila* bristles. BioEssays 18: 113–121

Maheshwari P, Singh H (1967) The female gametophyte of gymnosperms. Biol Rev 42:88–130

Millar CI (1983) A steep cline in *Pinus muricata*. Evolution 37: 311–319

Mouradov A, Glassick T, Hamdorf B, Murphy L, Fowler B, Maria S, Teasdale RD (1998) *Needly*, a *Pinus radiata* ortholog of *Floricaula*/*Leafy* genes, expressed in both reproductive and vegetative meristems. Proc Natl Acad Sci USA 95:6537–6542

Neale DB, Williams CG (1991) Restriction fragment length polymorphism mapping in conifers and applications to forest genetics and tree improvement. Can J For Res 21:545–554

Nelson CD, Nance WL, Doudrick RL (1993) A partial genetic-linkage map of slash pine (*Pinus elliottii* Engelm var 'Elliottii') based on random amplified polymorphic DNAs. Theor Appl Genet 87:145–151

Nelson CD, Kubisiak TL, Stine M, Nance WL (1994) A genetic-linkage map of longleaf pine (*Pinus palustris* Mill) based on random amplified polymorphic DNAs. J Hered 85:433–439

Plomion C, O'Malley DM, Durel CE (1995) Genomic analyses in maritime pine (Pinus pinaster) – comparison of 2 RAPD maps using selfed and open-pollinated seeds of the same individual. Theor Appl Genet 90:1028–1034

Plomion C, Durel CE, O'Malley DM (1996) Genetic dissection of height in maritime pine seedlings raised under accelerated growth conditions. Theor Appl Genet 93:849–858

Rogers DL (1997) Inheritance of allozymes from seed tissues of the hexaploid gymnosperm, *Sequoia sempervirens* (D. Don) Endl (Coast redwood). Heredity 78:166–175

Sederoff RR, Campbell MM, O'Malley DM, Whetten RW (1994) Genetic regulation of lignin biosynthesis and the potential modification of wood by genetic engineering in loblolly pine. Rec Adv Phytochem 28:313–355

Tulsieram LK, Glaubitz JC, Kiss G, Carlson JE (1992) Single tree genetic-linkage mapping in conifers using haploid DNA from megagametophytes. BioTechnology 10:686–690

Wang CT, Liu TP (1998) Inheritance and linkage relationships of allozymes, and estimation of outcrossing rates in a seed orchard of *Cunninghamia konishii* Hay. Silvae Genet 47:33–37

Wang ZM, Nagasaka K (1997) Allozyme variation in natural populations of *Picea glehnii* in Hokkaido, Japan. Heredity 78: 470–475

Wang ZM, Nagasaka K, Tanaka K (1996) Inheritance and linkage relationships of isozymes of *Picea glehnii* (Masters). Silvae Genet 45:136–141

Wheeler NC, Guries RP (1982) Population structure, genetic diversity, and morphological variation in *Pinus contorta* Dougl. Can J For Res 12:595–606

Wolfram, S (1996) The MATHEMATICA, 3rd edn. Wolfram Research, London

Wu RL (1998) Equations for estimating average effects and genetic variances of four megagametophyte genotypes at two epistatic-interacting genes. Technical Report, Department of Forestry, North Carolina State University, Raleigh, N.C.

Wu RL, O'Malley DM, Remington DL, Mackay JJ, McKeand SE (1999) Average effect of a mutation in lignin biosynthesis in loblolly pine. Theor Appl Genet 99:705–710

Wu RL (1999) Mapping quantitative trait loci by genotyping haploid tissues. Genetics 152:1741–1752

Zheng Y, Ennos (1997) Changes in the mating systems of populations of *Pinus caribaea* Morelet var 'caribaea' under domestication. Forest Genet 4:209–215